Johns Hopkins University, Dept. of Biostatistics Working Papers

12-15-2014

OPTIMAL, TWO STAGE, ADAPTIVE ENRICHMENT DESIGNS FOR RANDOMIZED TRIALS USING SPARSE LINEAR PROGRAMMING

Michael Rosenblum
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, mrosen@jhu.eduMi

Xingyuan Fang
Department of Operations Research and Financial Engineering, Princeton University

Han Liu
Department of Operations Research and Financial Engineering, Princeton University

Optimal, Two Stage, Adaptive Enrichment Designs for Randomized Trials, using Sparse Linear Programming

Michael Rosenblum*, Xingyuan Fang†, and Han Liu†

December 15, 2014

Abstract

Adaptive enrichment designs involve preplanned rules for modifying enrollment criteria based on accruing data in a randomized trial. Such designs have been proposed, for example, when the population of interest consists of biomarker positive and biomarker negative individuals. The goal is to learn which populations benefit from an experimental treatment. Two critical components of adaptive enrichment designs are the decision rule for modifying enrollment, and the multiple testing procedure. We provide the first general method for simultaneously optimizing both of these components for two stage, adaptive enrichment designs. We minimize expected sample size under constraints on power and the familywise Type I error rate. It is computationally infeasible to directly solve this optimization problem since it is not convex. The key to our approach is a novel representation of a discretized version of this optimization problem as a sparse linear program. We apply advanced optimization methods to solve this problem to high accuracy, revealing new, approximately optimal designs.

1 Introduction

Consider the problem of planning a randomized trial of a new treatment versus control, when the population of interest is partitioned into two subpopulations. The subpopulations could be defined in terms of a biomarker or risk score measured at baseline. Our goal is to test the

*mrosen@jhu.edu, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
†Department of Operations Research and Financial Engineering, Princeton University

null hypotheses of no average treatment benefit for each subpopulation and for the combined population. Standard randomized trial designs may have low power to detect a treatment effect if the treatment only benefits one subpopulation. Adaptive enrichment designs have been proposed for this problem, e.g., Follmann (1997), Russek-Cohen and Simon (1997), Jennison and Turnbull (2007), Wang et al. (2007), Wang et al. (2009), Brannath et al. (2009a), Rosenblum and van der Laan (2011), Jenkins et al. (2011a), Friede et al. (2012), Boessen et al. (2013a), Stallard et al. (2014).

An adaptive enrichment design consists of a decision rule for potentially modifying enrollment at an interim analysis, and a multiple testing procedure. The decision rule is allowed to be an arbitrary, prespecified function from the stage 1 data to a finite set of possible enrollment decisions for stage 2. The multiple testing procedure can be an arbitrary, prespecified function from the stage 1 and 2 data to the set of null hypotheses that are rejected. The class of possible designs is therefore quite large. Our goal is to construct new adaptive enrichment designs that minimize expected sample size under constraints on power and Type I error, over this class of possible designs. This is a nonconvex optimization problem that is computationally infeasible to solve directly.

Our approach is to approximate the original optimization problem by a large, sparse linear program. This idea was used in the context of standard (non-adaptive) designs by Rosenblum et al. (2014), where the only feature optimized was the multiple testing procedure. We tackle the substantially more challenging problem of simultaneously optimizing the decision rule and multiple testing procedure in two-stage, adaptive enrichment designs. The difficulty of the latter problem is twofold: it is both harder to construct a representation as a sparse, linear program, and the resulting linear program is harder to solve computationally. Another difference between the work here and Rosenblum et al. (2014) is that we consider not only power, but also expected sample size. In practice, both of these are important considerations in trial planning.

We prove that our designs control the familywise Type I error rate in the strong sense defined by Hochberg and Tamhane (1987, pg. 3). Control of the familywise Type I error rate is generally required by regulatory agencies such as the U.S. Food and Drug Administration for confirmatory randomized trials (FDA, 2010).

As in all of the above related work, we require the subpopulations to be defined before the trial starts. Such a definition could be based on prior trial data and scientific understanding of the disease being treated. Designs exist that try to solve the more challenging problem

of defining a subpopulation based on accruing data and then testing for a treatment effect in that subpopulation, e.g., Freidlin and Simon (2005); Lai et al. (2014). Optimizing trial designs in this context is an area of future work.

Hampson and Jennison (2014) consider the related problem of optimizing a two-stage adaptive design to determine the optimal treatment among k possible treatments, for a single population. Their general approach of converting the problem to a Bayes decision problem does not work in our setting, since this approach requires that the optimal solution at the global null hypothesis also controls familywise Type I error at all other alternatives. We show in Section 7 that this does not generally hold for our problem.

We focus on designs where the only allowed adaptation is to modify enrollment for stage 2. Other types of adaptive designs involve modifying randomization probabilities (called covariate-adaptive or response-adaptive designs), or modifying the treatment for each individual in response to his/her outcomes over time (called dynamic treatment regimes). In contrast to these types of adaptation, each participant in our designs is randomized with probability 1/2 to treatment or control, and once he/she is randomized there is no change to the treatment received. The only design feature that may be changed is the stage 2 enrollment criteria.

2 Problem Definition

2.1 Null Hypotheses

We assume the population is partitioned into two subpopulations, defined in terms of variables measured before randomization. Let p_s denote the proportion of the population in subpopulation $s \in \{1, 2\}$, which we assume are known; $p_1 + p_2 = 1$. Each enrolled participant is assigned to treatment ($a = 1$) or control ($a = 0$) with probability 1/2. Below, for clarity, we focus on normally distributed outcomes with known variances. However, we give asymptotic extensions of our results for a wide variety of outcome distributions in Section A of the Supplementary Materials.

For each subpopulation $s \in \{1, 2\}$ and stage $k \in \{1, 2\}$, we assume exactly half the participants are assigned to each study arm $a \in \{0, 1\}$. This can be approximately achieved in practice by using block randomization stratified by subpopulation. For each participant i from subpopulation $s \in \{1, 2\}$ enrolled in stage $k \in \{1, 2\}$, let $(A_{s,i}^{(k)}, Y_{s,i}^{(k)})$ denote his/her arm

assignment $A_{s,i}^{(k)} \in \{0,1\}$ and outcome $Y_{s,i}^{(k)} \in \mathbb{R}$, respectively. Throughout, the *sub*population indicator s is in the *sub*script, and the stage number k is in the superscript. We assume that conditioned on $A_{s,i}^{(k)} = a$, the outcome $Y_{s,i}^{(k)} \sim N(\mu_{sa}, \sigma_{sa}^2)$ and is independent of all other participant data. Let $\boldsymbol{\sigma}^2 = (\sigma_{10}^2, \sigma_{11}^2, \sigma_{20}^2, \sigma_{21}^2)$, which we assume are known. Let $X^{(k)}$ denote all the data from stage k, and let $X = X^{(1)} \cup X^{(2)}$ denote the cumulative data at the end of stage 2. Let $\mathcal{X}^{(k)}$ and $\mathcal{X}$ denote the sample spaces corresponding to $X^{(k)}$ and X, respectively.

Denote the population average treatment effect for each subpopulation $s \in \{1,2\}$ by $\Delta_s = \mu_{s1} - \mu_{s0}$, and for the combined population by $\Delta_C = p_1\Delta_1 + p_2\Delta_2$. Let $\boldsymbol{\Delta} = (\Delta_1, \Delta_2)$. Define H_{01}, H_{02}, H_{0C}, to be the null hypotheses of no average treatment benefit in subpopulation 1, subpopulation 2, and the combined population, respectively, i.e.,

$$H_{01} : \Delta_1 \leq 0; \qquad H_{02} : \Delta_2 \leq 0; \qquad H_{0C} : \Delta_C \leq 0.$$

Let $\mathcal{H} = \{H_{01}, H_{02}, H_{0C}\}$, and let $\mathcal{S}$ denote the power set of $\mathcal{H}$. For any $\boldsymbol{\Delta} \in \mathbb{R}^2$, define $\mathcal{H}_{\text{TRUE}}(\boldsymbol{\Delta})$ to be the set of true null hypotheses at $\boldsymbol{\Delta}$. For each $s \in \{1,2\}$, this set contains H_{0s} if $\Delta_s \leq 0$; it contains H_{0C} if $p_1\Delta_1 + p_2\Delta_2 \leq 0$.

2.2 Two-Stage Adaptive Enrichment Designs

In stage 1, $n_s^{(1)}$ participants are enrolled from each subpopulation s. At the interim analysis following stage 1, a decision rule D determines the number of participants to enroll from each subpopulation in stage 2. This decision is based on the data from stage 1, and there are $K < \infty$ possible decisions denoted by $\mathcal{D} = \{1, \ldots, K\}$. At the end of stage 2, a multiple testing procedure M determines which subset (if any) of the null hypotheses to reject, based on the data from stages 1 and 2. A two stage adaptive enrichment design is defined by the following quantities, which must be specified before the trial starts:

i. The stage 1 sample sizes $n_1^{(1)}, n_2^{(1)}$ for subpopulations 1 and 2, respectively.

ii. The number K of possible stage 2 decisions, and for each decision $d \in \mathcal{D} = \{1, \ldots, K\}$ the stage 2 sample sizes $n_1^{(2),d}, n_2^{(2),d}$ for subpopulations 1 and 2, respectively.

iii. A decision rule D mapping the stage 1 data $X^{(1)}$ to an enrollment decision in $\mathcal{D}$.

iv. A multiple testing procedure M mapping the stage 1 and 2 data X to a set of hypotheses $H \subseteq \mathcal{H}$ to reject.

Define an adaptive design template to be the quantities defined in (i)-(ii), i.e., the set of possible decisions and corresponding sample sizes $\mathbf{n} = (\mathcal{D}, n_1^{(1)}, n_2^{(1)}, \{n_1^{(2),d}, n_2^{(2),d}\}_{d\in\mathcal{D}})$. A general adaptive design template is displayed in Figure 1a. A specific example of an adaptive design template is given in Figure 1b for the case of $p_1 = 1/2$. In this example, for a given $n > 0$, the stage 1 sample sizes satisfy $n_1^{(1)} = n_2^{(1)} = n/4$. There are four choices for stage 2 enrollment: $D = 1$: stop the trial, i.e., $n_1^{(2),1} = n_2^{(2),1} = 0$; $D = 2$: enroll exactly as in stage 1, i.e., $n_1^{(2),2} = n_2^{(2),2} = n/4$; $D = 3$: only enroll from subpopulation 1, i.e., $n_1^{(2),3} = 3n/4, n_2^{(2),3} = 0$; $D = 4$: only enroll from subpopulation 2, i.e., $n_1^{(2),4} = 0, n_2^{(2),4} = 3n/4$.

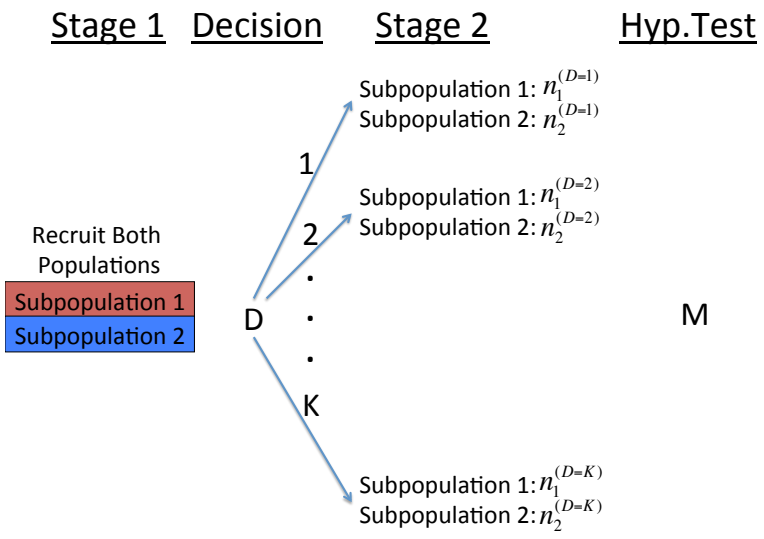
For a given adaptive design template $\mathbf{n}$, we aim to simultaneously optimize the decision rule D and multiple testing procedure M, in the sense defined in Section 2.3. The only constraints on D and M are that they are measurable functions as defined in Appendix B of the Supplementary Materials. Let $\mathcal{E}^*$ denote the class of all measurable functions from the sample space $\mathcal{X}^{(1)}$ to $\mathcal{D}$, and let $\mathcal{M}^*$ denote the class of all measurable functions from the sample space and decision $\mathcal{X} \times \mathcal{D}$ to the power set $\mathcal{S}$ of null hypotheses. For given values of $(\mathbf{n}, \boldsymbol{\sigma}, D, M)$, let $P_{\boldsymbol{\Delta}}$ denote the corresponding distribution of X and and let $E_{\boldsymbol{\Delta}}$ denote expectation with respect to this distribution.

2.3 General Optimization Problem

The quantity to be minimized, called the objective function, is defined in terms of a loss function L and a distribution Λ on the alternatives $\boldsymbol{\Delta}$. The loss function and distribution are set by the user to determine the quantity of interest to be optimized, e.g., these can be chosen to represent expected sample size and/or power as described below. We allow the loss function L to be any bounded, integrable function of the treatment effect $\boldsymbol{\Delta}$, the enrollment decision D, and the set of hypotheses rejected M. For a given loss function L, the risk at treatment effect vector $\boldsymbol{\Delta} \in \mathbb{R}^2$ is defined as $R_L(\boldsymbol{\Delta}) = E_{\boldsymbol{\Delta}} L[M\{X, D(X^{(1)})\}, D(X^{(1)}), \boldsymbol{\Delta}]$. The objective function is the Bayes risk $\int R_L(\boldsymbol{\Delta}) d\Lambda(\boldsymbol{\Delta})$. Below, for clarity of notation, we write D for $D(X^{(1)})$ and M for $M\{X, D(X^{(1)})\}$.

The above definition of the objective function allows us to select different quantities of interest (or weighted combinations of these quantities) to optimize, e.g., power, expected sample size, or expected number assigned to an ineffective treatment. For example, the loss function could be set equal to the total sample size $L^{\text{SS}} = n_1^{(1)} + n_2^{(1)} + n_1^{(2),D} + n_2^{(2),D}$; the corresponding risk at $\boldsymbol{\Delta} \in \mathbb{R}^2$ equals the expected sample size of the trial when the treatment

a.



b.

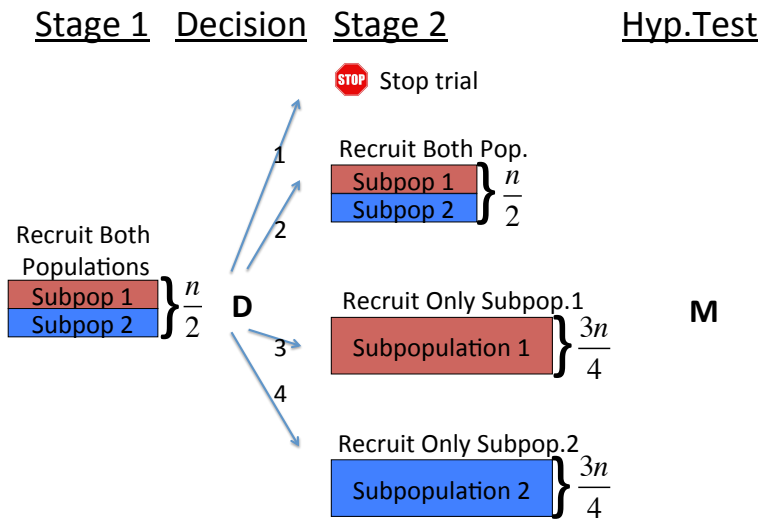


Figure 1: (a) Adaptive enrichment design template; (b) Example of adaptive enrichment design.

effect is $\boldsymbol{\Delta}$. Alternatively, we could encode power to reject different null hypotheses using the following loss functions:

$$\text{For each } s \in \{1,2\}, L^{(s)} = 1[H_{0s} \notin M; \Delta_s \geq \Delta^{\min}];$$
$$L^{(C)} = 1[H_{0C} \notin M, \Delta_1 \geq \Delta^{\min}, \Delta_2 \geq \Delta^{\min}],$$

where $\Delta^{\min}$ represents the minimum, clinically meaningful treatment effect, which is user-specified. The reason we put the constraint $\Delta_s \geq \Delta^{\min}$ in the loss function $L^{(s)}$ is that we only want to penalize for failing to reject H_{0s} when in truth the treatment effect for subpopulation s is above the clinically meaningful level. For each subpopulation $s \in \{1,2\}$, if the treatment effect Δ_s equals or exceeds the minimum level $\Delta^{\min}$, then the risk $R_{L^{(s)}}(\boldsymbol{\Delta})$ equals one minus the power to reject H_{0s}. Similarly, if both treatment effects Δ_1, Δ_2 equal or exceed the minimum level, the risk $R_{L^{(C)}}(\boldsymbol{\Delta})$ equals one minus the power to reject H_{0C}. In either case, minimizing risk corresponds to maximizing power.

We aim to minimize the Bayes risk, i.e., the risk integrated with respect to a distribution Λ on the alternatives $\boldsymbol{\Delta}$. For example, we could let Λ denote a weighted sum of the four point masses in the set $Q = \{(0,0), (\Delta^{\min}, 0), (0, \Delta^{\min}), (\Delta^{\min}, \Delta^{\min})\}$, which correspond to no treatment effect, only subpopulation 1 benefiting at the minimum level, only subpopulation 2 benefiting at the minimum level, and both subpopulations benefiting at the minimum level, respectively. Let Λ^{pm} denote this distribution with weight 1/4 on each point mass. Let Λ^{mix} denote a mixture of four normal distributions, with one centered at each of the aforementioned point masses, and each having variance σ^2_Λ. Then the Bayes risk corresponding to the pair $(L, \Lambda) = (L^{\text{SS}}, \Lambda^{\text{pm}})$ is the expected sample size under $\boldsymbol{\Delta}$, averaged over the four scenarios $\boldsymbol{\Delta} \in Q$. As another example, the Bayes risk corresponding to the pair $(L, \Lambda) = (a_1 L^{(1)} + a_2 L^{(2)}, \Lambda^{\text{mix}})$ for positive constants a_1, a_2 is the weighted sum of 1 minus the power to reject each H_{0s} when the corresponding treatment effect exceeds the minimum level, integrated over the distribution $\Lambda^{\text{mix}}(\boldsymbol{\Delta})$.

Our optimization problem has two types of constraints. The first are familywise Type I error constraints, and the second are additional constraints involving J triples $(L_j, \Lambda_j, \beta_j)$ of loss function L_j, distribution $\Lambda_j(\boldsymbol{\Delta})$, and threshold $\beta_j \in \mathbb{R}$ defined below.

Constrained Bayes Optimization Problem: For given $\mathbf{n}$, $\alpha > 0$, $\boldsymbol{\sigma}^2$, $\{(L_j, \Lambda_j, \beta_j) : j = 0, \ldots, J\}$, find the adaptive enrichment design $(D, M) \in (\mathcal{E}^* \times \mathcal{M}^*)$ minimizing

$$\int E_{\boldsymbol{\Delta}} \left(L_0[M\{X, D(X^{(1)})\}, D(X^{(1)}), \boldsymbol{\Delta}] \right) d\Lambda_0(\boldsymbol{\Delta}), \tag{1}$$

under the familywise Type I error constraints: for any $\boldsymbol{\Delta} \in \mathbb{R}^2$,

$$P_{\boldsymbol{\Delta}} \{M \text{ rejects any null hypotheses in } \mathcal{H}_{\text{TRUE}}(\boldsymbol{\Delta})\} \leq \alpha, \tag{2}$$

and additional constraints: for each $j \in \{1, \ldots, J\}$:

$$\int E_{\boldsymbol{\Delta}} \left(L_j[M\{X, D(X^{(1)})\}, D(X^{(1)}), \boldsymbol{\Delta}] \right) d\Lambda_j(\boldsymbol{\Delta}) \leq \beta_j. \tag{3}$$

First, consider the case where $J = 0$, i.e., there are no additional constraints (3). Then the constrained Bayes optimization problem is to minimize the Bayes risk subject to strong control on the familywise Type I error rate at level α. For example, one can optimize power in the sense described above by setting $(L, \Lambda) = (a_1 L^{(1)} + a_2 L^{(2)}, \Lambda^{\text{mix}})$.

2.4 Example of Optimization Problem

The additional constraints (3) allow the user to define a broader set of problems, such as optimizing expected sample size subject to power and Type I error constraints. For example, consider the problem of minimizing expected sample size averaged over the four point masses in Q, under the Type I error constraints (2) and the following power constraints for given Type II error $\beta > 0$:

P1. At $(\Delta_1, \Delta_2) = (\Delta^{\min}, 0)$, the power to reject H_{01} is at least $1 - \beta$.

P2. At $(\Delta_1, \Delta_2) = (0, \Delta^{\min})$, the power to reject H_{02} is at least $1 - \beta$.

P3. At $(\Delta_1, \Delta_2) = (\Delta^{\min}, \Delta^{\min})$, the power to reject H_{0C} is at least $1 - \beta$.

This problem can be represented by setting $(L, \Lambda) = (L^{\text{SS}}, \Lambda^{\text{pm}})$ and $J = 3$ additional constraints of the form $(L_j, \Lambda_j, \beta_j)$ equal to

$$(L^{(1)}, 1(\Delta^{\min}, 0), \beta); \qquad (L^{(2)}, 1(0, \Delta^{\min}), \beta); \qquad (L^{(C)}, 1(\Delta^{\min}, \Delta^{\min}), \beta),$$

where $1(x, y)$ denotes a point mass at $\boldsymbol{\Delta} = (x, y)$. We solve this problem in Section 5.

3 Reducing Problem Complexity through Minimal Sufficient Statistics

We show that it suffices to consider decision rules D and multiple testing procedures M that depend only on minimal sufficient statistics. This dramatically reduces the problem complexity from having to search over arbitrarily complex functions of the data X, to the easier (but still very challenging) problem of searching over functions of the 2-dimensional sufficient statistics at each stage. Let $N_s^{(k)}$ denote the number enrolled from subpopulation $s \in \{1,2\}$ during stage $k \in \{1,2\}$. The stage 1 sample sizes are set in advance, while the stage 2 sample sizes are functions of the stage 1 data; specifically, $N_s^{(1)} = n_s^{(1)}$ and $N_s^{(2)} = n_s^{(2),D(X^{(1)})}$ for each $s \in \{1,2\}$.

For each subpopulation $s \in \{1,2\}$ and stage $k \in \{1,2\}$, define the corresponding z-statistic as

$$Z_s^{(k)} = \left\{ \frac{\sum_{i=1}^{N_s^{(k)}} Y_{s,i}^{(k)} A_{s,i}^{(k)}}{\sum_{i=1}^{N_s^{(k)}} A_{s,i}^{(k)}} - \frac{\sum_{i=1}^{N_s^{(k)}} Y_{s,i}^{(k)} (1 - A_{s,i}^{(k)})}{\sum_{i=1}^{N_s^{(k)}} (1 - A_{s,i}^{(k)})} \right\} \left\{ \frac{\sigma_{s1}^2 + \sigma_{s0}^2}{N_s^{(k)}/2} \right\}^{-1/2}, \tag{4}$$

where the quantity inside curly braces on the right is the variance of the difference between sample means on the left. Let $\mathbf{Z}^{(k)} = (Z_1^{(k)}, Z_2^{(k)})$. Define the final (cumulative) z-statistic based on all stage 1 and 2 data for subpopulation s by

$$Z_s^{(F)} = \left\{ \frac{\sum_{k=1}^{2} \sum_{i=1}^{N_s^{(k)}} Y_{s,i}^{(k)} A_{s,i}^{(k)}}{\sum_{k=1}^{2} \sum_{i=1}^{N_s^{(k)}} A_{s,i}^{(k)}} - \frac{\sum_{k=1}^{2} \sum_{i=1}^{N_s^{(k)}} Y_{s,i}^{(k)} (1 - A_{s,i}^{(k)})}{\sum_{k=1}^{2} \sum_{i=1}^{N_s^{(k)}} (1 - A_{s,i}^{(k)})} \right\} \left\{ \frac{\sigma_{s1}^2 + \sigma_{s0}^2}{(N_s^{(1)} + N_s^{(2)})/2} \right\}^{-1/2}. \tag{5}$$

Let $\mathbf{Z}^{(F)} = (Z_1^{(F)}, Z_2^{(F)})$. The distribution of $\mathbf{Z} = (\mathbf{Z}^{(1)}, \mathbf{Z}^{(F)})$ is characterized as follows:

a. $\mathbf{Z}^{(1)}$ is bivariate normal with mean vector $\left(\Delta_1 \left\{ \frac{n_1^{(1)}}{2\left(\sigma_{11}^2 + \sigma_{10}^2\right)} \right\}^{1/2}, \Delta_2 \left\{ \frac{n_2^{(1)}}{2\left(\sigma_{21}^2 + \sigma_{20}^2\right)} \right\}^{1/2} \right)$ and covariance matrix I_2, i.e., the 2×2 identity matrix.

b. $\mathbf{Z}^{(2)}$, which uses only stage 2 data, is conditionally independent of $\mathbf{Z}^{(1)}$ given the decision $D(\mathbf{Z}^{(1)}, U_1)$. The conditional distribution of $\mathbf{Z}^{(2)}$ given $D(\mathbf{Z}^{(1)}, U_1) = d$ is bivariate normal with mean vector $\left(\Delta_1 \left\{ \frac{n_1^{(2),d}}{2\left(\sigma_{11}^2 + \sigma_{10}^2\right)} \right\}^{1/2}, \Delta_2 \left\{ \frac{n_2^{(2),d}}{2\left(\sigma_{21}^2 + \sigma_{20}^2\right)} \right\}^{1/2} \right)$ and covariance matrix I_2.

c. For each subpopulation $s \in \{1, 2\}$, for $D = D(\mathbf{Z}^{(1)}, U_1)$, we have the following relationship between the final (cumulative) z-statistic and the stagewise z-statistics:

$$Z_s^{(F)} = \left\{ \frac{n_s^{(1)}}{n_s^{(1)} + n_s^{(2),D}} \right\}^{1/2} Z_s^{(1)} + \left\{ \frac{n_s^{(2),D}}{n_s^{(1)} + n_s^{(2),D}} \right\}^{1/2} Z_s^{(2)}. \tag{6}$$

We show that it suffices to consider decision rules D that depend on the data only through $\mathbf{Z}^{(1)}$, and multiple testing procedures M that depend on the data only through $\mathbf{Z}^{(F)}$ and the decision D. We consider randomized decision rules and multiple testing procedures, i.e., we allow D and M to additionally take as input U_1 and U_2, respectively, which are independent, uniform random variables. For conciseness, we refer to "randomized decision rules" as "decision rules", and refer to "randomized multiple testing procedures" as "multiple testing procedures."

Let $\mathcal{E}$ denote the class of all measurable functions D from $\mathbb{R}^2 \times [0, 1]$ (representing all possible values of $\mathbf{Z}^{(1)}, U_1$) to the set of stage 2 enrollment decisions $\mathcal{D}$. Let $\mathcal{M}$ denote the class of all measurable functions from $\mathbb{R}^2 \times \mathcal{D} \times [0, 1]$ (representing all possible values of $\mathbf{Z}^{(F)}, D(\mathbf{Z}^{(1)}, U_1), U_2$) to $\mathcal{S}$ (indicating the subset of null hypotheses rejected). For conciseness, we let $D = D(\mathbf{Z}^{(1)}, U_1)$ and $M = M\{\mathbf{Z}^{(F)}, D(\mathbf{Z}^{(1)}, U_1), U_2\}$. Define the class of adaptive enrichment designs $\mathcal{A} = \{(D, M) : D \in \mathcal{E}, M \in \mathcal{M}\}$. We prove the following theorem in Section J of the Supplementary Materials:

Theorem 3.1. *If the constrained Bayes optimization problem in Section 2.3 is feasible, then there exists an optimal solution $(D, M) \in (\mathcal{E} \times \mathcal{M})$, i.e., for which D depends on the data only through $\mathbf{Z}^{(1)}$, and M depends on the data only through $\mathbf{Z}^{(F)}$ and the decision $D(\mathbf{Z}^{(1)}, U_1)$.*

4 Discretization of Constrained Bayes Optimization Problem and Transformation into Sparse Linear Program

4.1 Discretization of Constrained Bayes Optimization Problem

Even after simplifying the constrained Bayes optimization problem by using only minimal sufficient statistics as in the previous section, the problem is still extremely difficult or impossible to solve directly. This is because the optimization problem is nonconvex, involves infinitely many familywise Type I error constraints (2), and optimizes over the very large

class of decision rules $\mathcal{E}$ and multiple testing procedures $\mathcal{M}$. We propose a novel approach to solve this problem, involving three steps. We first discretize the decision rule and multiple testing procedure, and restrict to a finite subset of familywise Type I error constraints. The resulting problem is still nonconvex, and so is extremely difficult to solve. Step two involves reparametrizing this problem so that it can be represented as a sparse, linear program, a class of problems that is much easier to solve than nonconvex problems. The third step is to apply advanced optimization methods to solve the sparse, linear program.

The first of the above steps is to discretize the constrained Bayes optimization problem. The decision rule D is discretized by partitioning $\mathbb{R}^2$ into a finite set of rectangles as described below. The intuition for what follows is that we restrict to the subclass of adaptive designs $(D, M) \in (\mathcal{E} \times \mathcal{M})$ such that the following hold: D makes the same decision when the first stage statistics $\mathbf{Z}^{(1)}$ are anywhere within a small rectangle $r \subseteq \mathbb{R}^2$; M rejects the same set of null hypotheses when the first stage statistics $\mathbf{Z}^{(1)}$ are in a rectangle $r \subseteq \mathbb{R}^2$, the enrollment decision is $d \in \mathcal{D}$, and the final statistics $\mathbf{Z}^{(F)}$ are in a rectangle r'. For a fine enough partition of rectangles, we expect the solution to the corresponding discretized optimization problem to be close to that of the original problem.

We consider partitions of $\mathbb{R}^2$ into rectangles. One way to construct such a partition is to start with a box $B = [-b, b] \times [-b, b]$, for a given integer $b > 0$. Let $\tau = (\tau_1, \tau_2)$ be such that b/τ_s is an integer for each $s \in \{1, 2\}$. For each $k, k' \in \mathbb{Z}$, define the rectangle $R_{k,k'} = [k\tau_1, (k+1)\tau_1) \times [k'\tau_2, (k'+1)\tau_2)$. Let $\mathcal{R}_{\mathcal{B}}$ denote the set of such rectangles in the bounded region B, i.e., $\{R_{k,k'} : k, k' \in \mathbb{Z}, R_{k,k'} \subset B\}$. Define the following partition of $\mathbb{R}^2$: $\mathcal{R} = \mathcal{R}_{\mathcal{B}} \cup \{\mathbb{R}^2 \setminus B\}$. Though $\mathbb{R}^2 \setminus B$ is not a rectangle, we still refer to $\mathcal{R}$ as a partition of rectangles, with a slight abuse of notation.

Let $\mathcal{R}_{\text{dec}}$ denote a partition of $\mathbb{R}^2$ into rectangles. We restrict to the subclass of decision rules D with the following property: for any rectangle $r \in \mathcal{R}_{\text{dec}}$ and $u \in [0, 1]$,

$$D(\mathbf{z}^{(1)}, u) = D(\mathbf{z}^{(1)'}, u) \text{ whenever } \mathbf{z}^{(1)} \text{ and } \mathbf{z}^{(1)'} \text{ are both in } r. \tag{7}$$

That is, the decision rule only depends on the data through the rectangle that the first stage z-statistics are in.

For each $d \in \mathcal{D}$, let $\mathcal{R}_{\text{mtp},d}$ denote a partition of $\mathbb{R}^2$ into rectangles. Intuitively, we will restrict to multiple testing procedures M that only depend on the data through the enrollment decision D and the rectangles that the first stage and cumulative statistics are in, respectively. Let $\mathbf{z} = (\mathbf{z}^{(1)}, \mathbf{z}^{(F)})$ and $\mathbf{z}' = (\mathbf{z}^{(1)'}, \mathbf{z}^{(F)'})$, for any $\mathbf{z}^{(1)}, \mathbf{z}^{(F)}, \mathbf{z}^{(1)'}, \mathbf{z}^{(F)'} \in \mathbb{R}^2$.

We restrict to the subclass of multiple testing procedures M such that for any $r \in \mathcal{R}_{\text{dec}}$, $d \in \mathcal{D}$, $r' \in \mathcal{R}_{\text{mtp},d}$, and $u_1, u_2 \in [0,1]$, we have $M(\mathbf{z}, d, u_2) = M(\mathbf{z}', d, u_2)$ whenever all of the following hold: $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(1)'}$ are both in r, $D(\mathbf{z}^{(1)}, u_1) = D(\mathbf{z}^{(1)'}, u_1)$, and $\mathbf{z}^C$ and $\mathbf{z}^{C'}$ are both in r'.

For each $r \in \mathcal{R}_{\text{dec}}$ and $d \in \mathcal{D}$, define x_{rd} to be the probability that decision d is made conditioned on $\mathbf{Z}^{(1)} \in r$, i.e.,

$$x_{rd} = P\left\{D(\mathbf{Z}^{(1)}, U_1) = d | \mathbf{Z}^{(1)} \in r\right\}. \tag{8}$$

For each $r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}$, define $y_{rdr's}$ to be the probability that precisely the subset s is rejected conditioned on $\mathbf{Z}^{(1)} \in r$, $D(\mathbf{Z}^{(1)}, U_1) = d, \mathbf{Z}^{(F)} \in r'$, i.e.,

$$y_{rdr's} = P\left\{M(\mathbf{Z}, d, U_2) = s | \mathbf{Z}^{(1)} \in r, D(\mathbf{Z}^{(1)}, U_1) = d, \mathbf{Z}^{(F)} \in r'\right\}. \tag{9}$$

The values of all variables x_{rd} and $y_{rdr's}$ are specified by the study designer before the trial, and our goal is to optimize the corresponding Bayes risk under familywise Type I error constraints and the additional constraints (3).

The probability of rejecting precisely the subset $s \in \mathcal{S}$ at a given vector of population parameters $\boldsymbol{\Delta} = (\Delta_1, \Delta_2) \in \mathbb{R}^2$ is

$$\begin{aligned}
& P_{\boldsymbol{\Delta}}\left\{M(\mathbf{Z}, D(\mathbf{Z}^{(1)}, U_1), U_2) = s\right\} && (10)\\
= & \sum_{r,d,r'} P_{\boldsymbol{\Delta}}\left\{\mathbf{Z}^{(1)} \in r, D(\mathbf{Z}^{(1)}, U_1) = d, \mathbf{Z}^{(F)} \in r', M(\mathbf{Z}, d, U_2) = s\right\} \\
= & \sum_{r,d,r'} \left[P_{\boldsymbol{\Delta}}\left\{M(\mathbf{Z}, d, U_2) = s | \mathbf{Z}^{(F)} \in r', \mathbf{Z}^{(1)} \in r, D(\mathbf{Z}^{(1)}, U_1) = d\right\} \times \right. \\
& \left. P_{\boldsymbol{\Delta}}\left\{\mathbf{Z}^{(F)} \in r' | \mathbf{Z}^{(1)} \in r, D(\mathbf{Z}^{(1)}, U_1) = d\right\} P_{\boldsymbol{\Delta}}\left\{D(\mathbf{Z}^{(1)}, U_1) = d | \mathbf{Z}^{(1)} \in r\right\} P_{\boldsymbol{\Delta}}\left\{\mathbf{Z}^{(1)} \in r\right\}\right] \\
= & \sum_{r,d,r'} x_{rd} y_{rdr's} p(\boldsymbol{\Delta}, r, d, r'), && (11)
\end{aligned}$$

where (11) follows from (8) and (9), and where we define

$$p(\boldsymbol{\Delta}, r, d, r') = P_{\boldsymbol{\Delta}}\left\{\mathbf{Z}^{(F)} \in r' | \mathbf{Z}^{(1)} \in r, D(\mathbf{Z}^{(1)}, U_1) = d\right\} P_{\boldsymbol{\Delta}}\left\{\mathbf{Z}^{(1)} \in r\right\}. \tag{12}$$

The value of $p(\boldsymbol{\Delta}, r, d, r')$ does not depend on D, which follows from (7). This value can be computed to high precision using the multivariate normal distribution function and (a)-(c) from Section 2.2, as described in Section F of the Supplementary Material.

We can express the objective function (1) of the constrained Bayes optimization problem in terms of the variables $x_{rd}y_{rdr's}$, since the expectation inside the integral in (1) satisfies

$$\begin{aligned} E_{\boldsymbol{\Delta}}\left\{L(M(\mathbf{Z}, D(\mathbf{Z}^{(1)}, U_1), U_2); \Delta_1, \Delta_2)\right\} &= \sum_{s\in\mathcal{S}} L(s;\Delta_1,\Delta_2) P(M(\mathbf{Z}, D(\mathbf{Z}^{(1)}, U_1), U_2) = s) \\ &= \sum_{s\in\mathcal{S}} L(s;\Delta_1,\Delta_2) \sum_{r,d,r'} x_{rd}y_{rdr's} p(\boldsymbol{\Delta}, r, d, r') \\ &= \sum_{s\in\mathcal{S}} \sum_{r,d,r'} x_{rd}y_{rdr's} \left\{L(s;\Delta_1,\Delta_2) p(\boldsymbol{\Delta}, r, d, r')\right\}. \end{aligned}$$

where the second line follows from the equality of (10) and (11). The familywise Type I error constraints and additional constraints (3) can similarly be expressed as a function of $x_{rd}y_{rdr's}$, as shown in Section G of the Supplementary Materials.

Let $G \subset \mathbb{R}^2$ denote a discretization of the boundaries of the null spaces of the hypotheses of interest; an example is given in Section 7. The discretized version of the Constrained Bayes Optimization Problem above is as follows:

Discretized Problem:

$$\min \sum_{r,d,r',s} x_{rd}y_{rdr's} \int L_0(s;\boldsymbol{\Delta}) p(\boldsymbol{\Delta}, r, d, r') d\Lambda_0(\boldsymbol{\Delta}) \tag{13}$$

under the following constraints:

$$\text{for each } \boldsymbol{\Delta} \in G, \sum_{r,d,r'} \sum_{s\in\mathcal{S}: s\cap\mathcal{H}_{TRUE}(\boldsymbol{\Delta})\neq\emptyset} x_{rd}y_{rdr's} p(\boldsymbol{\Delta}, r, d, r') \leq \alpha; \tag{14}$$

$$\text{for each } j \in \{1,\ldots,J\}, \sum_{r,d,r',s} x_{rd}y_{rdr's} \int L_j(s;\boldsymbol{\Delta}) p(\boldsymbol{\Delta}, r, d, r') d\Lambda_j(\boldsymbol{\Delta}) \leq \beta_j; \tag{15}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, \sum_{d\in\mathcal{D}} x_{rd} = 1; \tag{16}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, \sum_{s\in\mathcal{S}} y_{rdr's} = 1; \tag{17}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S} : x_{rd} \geq 0, y_{rdr's} \geq 0. \tag{18}$$

The sum $\sum_{r,d,r',s}$, which appers in (13) and (15), is taken over $r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}$. The objective function (13) represents (1). The constraints (14) and (15) represent the familywise Type I error constraints (2) and additional constraints (3), respectively. The remaining constraints encode properties of x_{rd} and $y_{rdr's}$ that follow from their definitions (8)-(9) as conditional probabilities. Specifically, the constraints (16) and (17) follow from the law of total probability; the constraints (18) encode that each variable must be nonnegative since it represents a probability.

4.2 Transformation of (Nonconvex) Discretized Problem into Sparse Linear Program

The discretized problem from Section 4 is not linear (and not convex) in the variables $\{x_{rd}, y_{rdr's}\}$. Therefore, this problem is generally computationally intractable to solve, since only ad hoc methods exist for solving nonconvex optimization problems and even if a local minimum is found there is no general way to determine if it is the global minimum. We transform this problem into a sparse, linear program by defining the new variables:

$$v_{rdr's} = x_{rd} y_{rdr's}, \text{ for all } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}. \tag{19}$$

The objective function (13) and familywise Type I error constraints (14) are linear functions of $v_{rdr's}$. We prove in Section H of the Supplementary Material that the constraints (16)-(18) can be equivalently expressed in terms of the linear constraints (23)-(25) on $v_{rdr's}$ in the following linear program, where for each $d \in \mathcal{D}$, we let r'_d an arbitrary element in the set $\mathcal{R}_{\text{mtp},d}$ (say, the first element under a fixed ordering of $\mathcal{R}_{\text{mtp},d}$):

Sparse linear program:

$$\min \sum_{r,d,r',s} v_{rdr's} \int L_0(s;\boldsymbol{\Delta}) p(\boldsymbol{\Delta}, r, d, r') d\Lambda_0(\boldsymbol{\Delta}) \tag{20}$$

under the constraints:

$$\text{for each } \boldsymbol{\Delta} \in G, \sum_{s \in \mathcal{S}: s \cap \mathcal{H}_{TRUE}(\boldsymbol{\Delta}) \neq \emptyset} \sum_{r,d,r'} v_{rdr's} p(\boldsymbol{\Delta}, r, d, r') \leq \alpha; \tag{21}$$

$$\text{for each } j \in \{1, \ldots, J\}, \sum_{r,d,r',s} v_{rdr's} \int L_j(s;\boldsymbol{\Delta}) p(\boldsymbol{\Delta}, r, d, r') d\Lambda_j(\boldsymbol{\Delta}) \leq \beta_j; \tag{22}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, \sum_{d} \sum_{s \in \mathcal{S}} v_{rdr'_d s} = 1; \tag{23}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, \widetilde{r}' \in \mathcal{R}_{\text{mtp},d}, \sum_{s \in \mathcal{S}} v_{rdr'_d s} = \sum_{s \in \mathcal{S}} v_{rd\widetilde{r}'s}; \tag{24}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S} : v_{rdr's} \geq 0. \tag{25}$$

We prove the following theorem in Section I of the Supplementary Material:

Theorem 4.1. *i. (Equivalence of discretized problem and sparse linear program) The optimum value of the above optimization problem equals the optimum value of the discretized problem from Section 4.1.*

ii. (Map from solution of sparse linear program to solution of discretized problem) For any optimal solution $\mathbf{v} = \{v_{rdr's}\}_{r\in\mathcal{R}_{dec},d\in\mathcal{D},r'\in\mathcal{R}_{mtp,d},s\in\mathcal{S}}$ *to the sparse linear program, define the vectors* $\mathbf{x}, \mathbf{y}$ *by the transformation:*

$$x_{rd} = \sum_{s\in\mathcal{S}} v_{rdr's}; \tag{26}$$

$$y_{rdr's} = \begin{cases} v_{rdr's}/x_{rd}, & \text{if } x_{rd} > 0 \\ 1/|\mathcal{S}|, & \text{otherwise} \end{cases} . \tag{27}$$

Then $\mathbf{x}, \mathbf{y}$ *is a well-defined, feasible, and optimal solution to the discretized problem from Section 4.1.*

5 Application to Minimizing Expected Sample Size under Power and Type I Error Constraints

We consider the adaptive design template in Figure 1b from Section 2.2 and the optimization problem in Section 2.4. Let ESS_Q denote the value of the objective function (1), which equals the expected sample size averaged over the four mass points in Q. The sample sizes $\mathbf{n}$ are a function of n as described in Section 2.2, where n is the total sample size if both subpopulations are enrolled during stage 2. Let $p_1 = 1/2$, $\alpha = 0.05$, and let each σ^2_{sa} equal a common value $\sigma^2 > 0$. Let Φ denote the cumulative distribution function of the standard normal. For comparison purposes, for $\widetilde{\beta} = 0.05$, we set

$$n = 4\sigma^2\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\widetilde{\beta})\}^2/(\Delta^{\min})^2, \tag{28}$$

i.e., the smallest n such that in a standard (non-adaptive) design enrolling $n/2$ from each subpopulation, the uniformly most powerful test of H_{0C} at level $\alpha = 0.05$ has power $1-\widetilde{\beta} = 0.95$ at the alternative $\boldsymbol{\Delta} = (\Delta^{\min}, \Delta^{\min})$.

The optimal solution to the above constrained Bayes optimization problem is the same regardless of the choice of $(\sigma^2, \Delta^{\min})$, as we prove in Section P of the Supplementary Material. In brief, the reason is that by (a)-(c) in Section 3, both the distribution of $\mathbf{Z}^{(1)}$ and the conditional distribution of $\mathbf{Z}^{(2)}$ given $(D = d, \mathbf{Z}^{(1)})$ depend on $(n, \sigma^2, \boldsymbol{\Delta})$ only through $\Delta_1\{n/(8\sigma^2)\}^{1/2}$ and $\Delta_2\{n/(8\sigma^2)\}^{1/2}$, i.e., the non-centrality parameters for the subpopulations. For each $j \in \{0,1,2,3\}$, the support of the distribution Λ_j is contained

in $Q = \{(0,0),(\Delta^{\min},0),(0,\Delta^{\min}),(\Delta^{\min},\Delta^{\min})\}$. Therefore, the probabilities in the objective function (1) and additional constraints (3) depend on $(n,\sigma^2,\Delta^{\min})$ only through $\Delta^{\min}\{n/(8\sigma^2)\}^{1/2}$, which equals the constant $2^{-1/2}\{\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\widetilde{\beta})\}$ by (28).

We applied the method from Section 4.2 to solve the above problem for each $\beta \in \{0.01,\ldots,0.99\}$. Our results show the problem is feasible whenever $1-\beta \leq 0.82$, and is infeasible otherwise. We focus on the case of $1-\beta = 0.82$, and denote the corresponding decision rule and multiple testing procedure for the optimal solution by D^* and M^*, respectively. These are depicted in Figure 2.

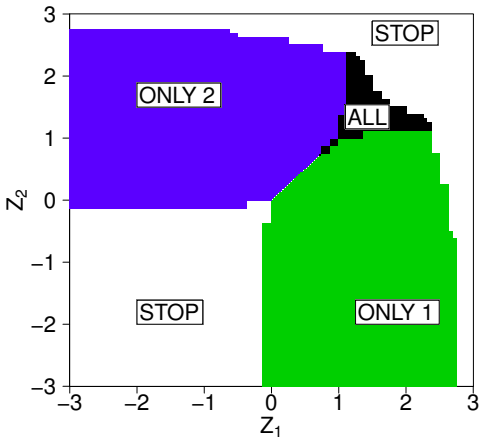
Let D^{std} denote the decision rule corresponding to a standard (non-adaptive) design that always enrolls from both subpopulations in stage 2, i.e., $D^{\text{std}} = 2$ for all values of the stage 1 data. Consider the above constrained Bayes optimization problem if we restrict to the standard design, i.e., all $(D,M) \in (\mathcal{E} \times \mathcal{M})$ that satisfy $D = D^{\text{std}}$. Then the problem above is infeasible for any value of $1-\beta > 0.65$, which we verified using the method from Section 4.2. In contrast, the adaptive enrichment design is feasible when $1-\beta \leq 0.82$, i.e., it can achieve 82% power for all three cases (P1)-(P3), while the standard design using the optimal multiple testing procedure in $\mathcal{M}$ can achieve power at most 0.64 for all three of these cases. We similarly considered the above optimization problem restricted to a standard design with total sample size $5n/4$, i.e., the maximum total sample size that can occur in the adaptive enrichment design; this problem is infeasible for any standard design when $1-\beta > 0.73$. So there is still a substantial advantage for the adaptive enrichment design.

6 Comparison of Optimal Adaptive Enrichment Design Versus Design Based on P-value Combination Approach

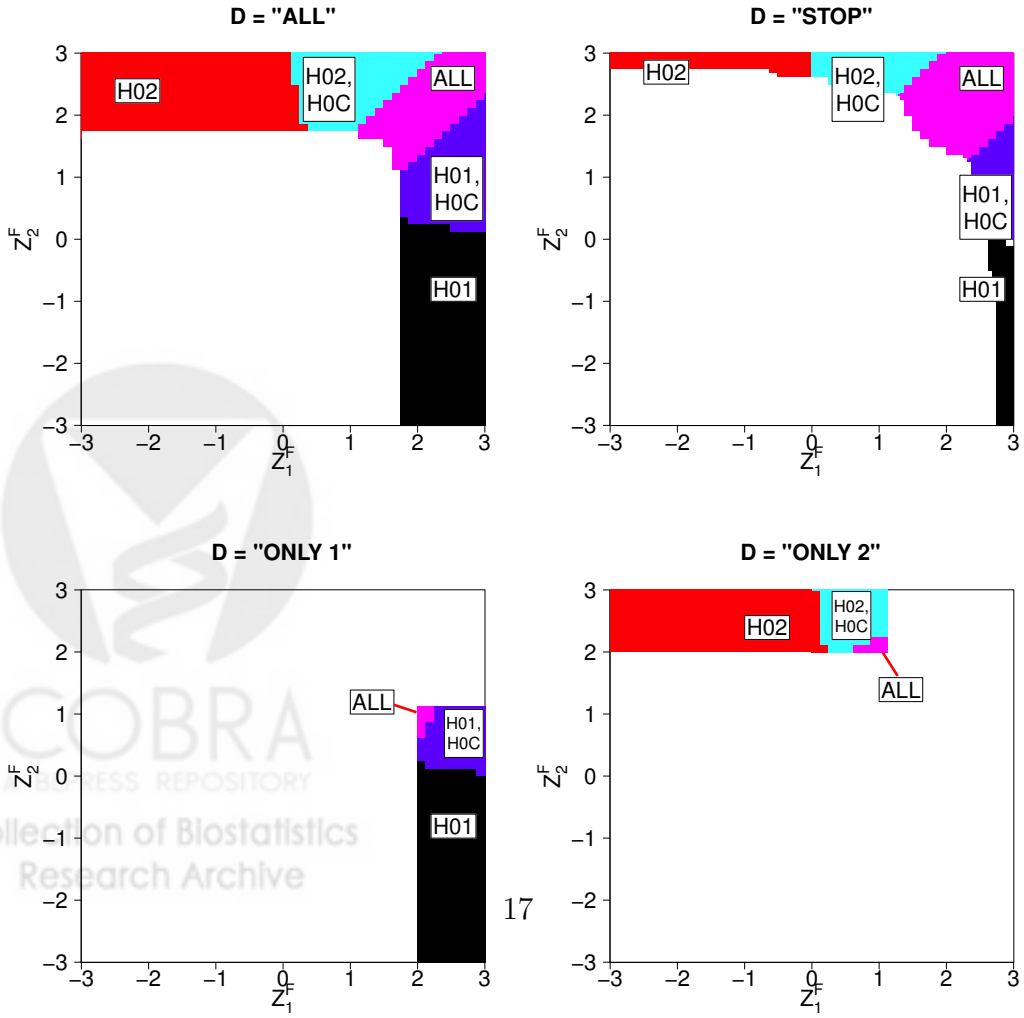
Consider the adaptive design template in Figure 1b from Section 2.2 and the optimization problem in Section 2.4. We apply the p-value combination approach of Bauer (1989), Bauer and Köhne (1994), Lehmacher and Wassmer (1999), with the closed testing principle of Marcus et al. (1976); this approach has been used to construct adaptive enrichment designs by, e.g., Bretz et al. (2006); Schmidli et al. (2006); Jennison and Turnbull (2007); Brannath et al. (2009b); Jenkins et al. (2011b); Boessen et al. (2013b). Multiple testing procedures M based on this approach are flexible in that they strongly control the familywise Type I error

Figure 2: Optimal Decision rule D^* and Multiple Testing Procedure M^* for Adaptive Enrichment Design Solving Optimization Problem in Section 5

Decision Rule D^* for Stage 2 Enrollment (z-statistics correspond to $\mathbf{Z}^{(1)}$):



Rejection Regions of M^* Corresponding to Each Possible Decision:



rate regardless of what decision rule D is used.

We applied the p-value combination approach to determine the multiple testing procedure M. This approach requires specifying a combination function and local tests for each intersection of null hypotheses. At each stage $k \in \{1, 2\}$, for every subset of null hypotheses $I \subseteq \mathcal{H}$, an adjusted p-value for the intersection null hypothesis $H_I = \bigcap_{H \in I} H$ is computed from the data in stage k, using the Dunnett intersection test (Dunnett, 1955; Jennison and Turnbull, 2007). P-values are then combined across stages by the weighted inverse normal rule with equal weights for each stage. Each elementary null hypothesis $H \in \mathcal{H}$ is rejected if and only if the stage 2 combined p-value is less than 0.05 for every intersection null hypothesis $\bigcap_{H' \in I} H'$ for which $H \in I$. We slightly modified this approach to incorporate early stopping for efficacy after stage 1 as in, e.g., Jennison and Turnbull (2007), using the equivalent of the boundaries of O'Brien and Fleming (1979) for the stage 1 p-values. Full details of our implementation are given in Section Q of the Supplementary Material. The resulting multiple testing procedure is denoted by M^{pv}.

The p-value combination approach does not specify a corresponding decision rule D. To construct one, we consider a class of decision rules D that are functions of two thresholds t_c and t_i, which we approximately optimize in conjunction with M^{pv} as described below. Sample sizes $\mathbf{n}$ are as in the adaptive design template in Figure 1b from Section 2.2. Define the decision rule $D^{(t_c,t_i)}(\mathbf{Z}^{(1)})$ as follows: If the multiple testing procedure M^{pv} rejects any null hypothesis at the end of stage 1, stop the trial; else, if the combined population statistic $(Z_1^{(1)} + Z_2^{(1)})/\sqrt{2} > t_c$, enroll both subpopulations in stage 2; else, enroll from each subpopulation s for which $Z_s^{(1)} > t_i$. The numerical value of D is then determined as follows: if both subpopulations are enrolled in stage 2, then $D = 2$; else, if only one subpopulation $s \in \{1, 2\}$ is enrolled in stage 2, then $D = 2 + s$; else, the trial stops at the end of stage 1, i.e., $D = 1$. An example of the decision rule $D^{(t_c,t_i)}$ is depicted in Figure 6.

We next define a set of adaptive enrichment designs $D^{(t_c,t_i)}$ corresponding to pairs (t_c, t_i) in a grid of values; let $\mathcal{A} = \{(D^{(t_c,t_i)}, M^{\text{pv}}) : (t_c, t_i) \in (-3, -2.9, \ldots, 3) \times (-3, -2.9, \ldots, 3)\}$. Each design in $\mathcal{A}$ strongly controls the familywise Type I error rate at level 0.05, which is a property of the p-value combination approach. For each design in $\mathcal{A}$, we computed ESS_Q and the power to reject each subset of null hypotheses. We used the results to solve the constrained Bayes optimization problem from the previous section restricted to the set of designs $\mathcal{A}$. Specifically, for each value of $1 - \beta$ in the top row of Table 1, we computed the smallest value of ESS_Q over all designs $\mathcal{A}$ that satisfy (P1)-(P3) at this value of β.

Figure 3: Decision Rule $D^{(t_c,t_i)}$ for $(t_c, t_i) = (1.6, 0.6)$. (z-statistics correspond to $\mathbf{Z}^{(1)}$). This corresponds to the minimizer of ESS_Q over $\mathcal{A}$ satisfying the power constraints (P1)-(P3) at $1 - \beta = 0.74$. The white area in the upper right corner corresponds to stopping the trial at the end of stage 1.

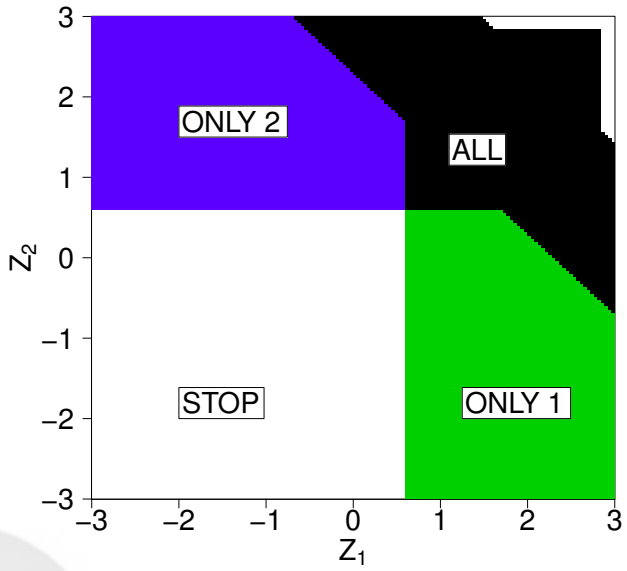


Table 1: The minimum value of ESS_Q, among the designs $\mathcal{A}$ (computed using grid search and p-value combination approach) and among the designs $\mathcal{E} \times \mathcal{M}$ (computed using sparse linear programming approach), for various values $1-\beta$ of the power constraints (P1)-(P3). No value is given for $\mathcal{A}$ when $1-\beta \geq 0.78$ since the problem is infeasible.

Power Constraint $(1-\beta)$	58%	62%	66%	70%	74%	78%	82%
Min. ESS_Q over $\mathcal{A}$	$0.86n$	$0.89n$	$0.92n$	$0.96n$	$1.02n$	$\times$	$\times$
Min. $ESS+Q$ over $\mathcal{E} \times \mathcal{M}$	$0.65n$	$0.69n$	$0.73n$	$0.79n$	$0.84n$	$0.92n$	$1.03n$

Also, for each such value of $1-\beta$, we solved the corresponding problem over the class of designs $\mathcal{E} \times \mathcal{M}$ using the sparse linear programming method from Section 4.2; the minimum value of the objective function ESS_Q is given for each class of designs in the bottom rows of Table 1 in terms of n. At all values of $1-\beta$ we considered, the minimum value of ESS_Q was substantially lower for the optimal design among $\mathcal{E} \times \mathcal{M}$ computed based on our sparse linear programming approach, compared to the optimal design among $\mathcal{A}$ computed using grid search and p-value combination approach. E.g., at $1-\beta = 0.74$, the value of ESS_Q for the former is 21% smaller than for the latter. In addition, the optimization problem is infeasible for the designs in $\mathcal{A}$ at $1-\beta \geq 0.78$, i.e., it is not possible to simultaneously satisfy the power constraints (P1)-(P3); in contrast, the problem is feasible for the class $\mathcal{E} \times \mathcal{M}$ up to power threshold $1-\beta = 0.82$. This shows that there are substantial gains in expected sample size and power from using the optimal design over the class $\mathcal{E} \times \mathcal{M}$ compared to the optimal design over $\mathcal{A}$. Our sparse linear programming method made it possible to compute the optimal design over $\mathcal{E} \times \mathcal{M}$, which previously was an open problem.

7 General Form of Sparse Linear Program

We describe the general form of the sparse linear program from Section 4.2. Let $\mathbf{v}$ denote the vector consisting of all variables $v_{rdr's}$ for $r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}$ in lexicographic order. Let $w = |\mathcal{R}_{\text{dec}}| \times \sum_{d \in \mathcal{D}} |\mathcal{R}_{\text{mtp},d}| \times |\mathcal{S}|$, where $|B|$ denotes the cardinality of the set B. Then $\mathbf{v}$ has w components. Define $\psi = \sum_{d \in \mathcal{D}} |\mathcal{R}_{\text{mtp},d}|$. The general form of the sparse linear program from Section 4.2 is

$$\min_{\mathbf{v}\in\mathbb{R}_+^w} \quad \mathbf{c^T v} \qquad \text{s.t.} \quad \mathbf{A^{(1)}v} \leq \mathbf{a^{(1)}}, \mathbf{A^{(2)}v} = \mathbf{a^{(2)}}; \tag{29}$$

for matrices $\mathbf{A^{(1)}}, \mathbf{A^{(2)}}$ and vectors $\mathbf{c}, \mathbf{a^{(1)}}, \mathbf{a^{(2)}}$. The matrix $\mathbf{A^{(1)}}$ has dimensions $(|G|+J)\times w$, and the matrix $\mathbf{A^{(2)}}$ has dimensions $(1+\psi)|\mathcal{R}_{\text{dec}}|\times w$. The matrix $\mathbf{A^{(1)}}$ is dense (most entries are non-zero), while the matrix $\mathbf{A^{(2)}}$ is sparse (most entries are 0) and has the form:

$$\mathbf{A^{(2)}} = \left[\begin{array}{c} |\mathcal{R}_{\text{dec}}| \text{ rows, each with } |\mathcal{D}| \times |\mathcal{S}| \text{ entries with 1 and the rest 0's.} \\ \hline \psi|\mathcal{R}_{\text{dec}}| \text{ rows, each with } |\mathcal{S}| \text{ entries = 1, } |\mathcal{S}| \text{ entries = -1, and the rest 0's.} \end{array}\right].$$

Though the matrix $\mathbf{A^{(2)}}$ is typically much larger than $\mathbf{A^{(1)}}$, the former does not dramatically impact the computationally difficulty since it is sparse.

The vector $\mathbf{c}$ is dense, and the vectors $\mathbf{a^{(1)}}, \mathbf{a^{(2)}}$ have the following forms:

$$\mathbf{a^{(1)}}^T = \begin{pmatrix} |G| \text{ entries} & J \text{ entries} \\ \alpha, \ldots, \alpha, & \beta_1, \ldots, \beta_J \end{pmatrix}, \ \mathbf{a^{(2)}}^T = \begin{pmatrix} |\mathcal{R}_{\text{dec}}| \text{ entries} & \psi|\mathcal{R}_{\text{dec}}| \text{ entries} \\ 1, \ldots 1, & 0, \ldots, 0 \end{pmatrix}.$$

Consider the constrained Bayes optimization problem in Section 5. The problem was solved in two steps. First, a sparse linear program was constructed using a coarse discretization with $w = 479,570$ variables, 541 Type I error constraints corresponding to pairs of non-centrality parameters $(\Delta_1\{n/(8\sigma^2)\}^{1/2}, \Delta_2\{n/(8\sigma^2)\}^{1/2})$ in the set $\{(x,y) : [x \in \{-9,-8.9,\ldots,9\}, y = 0]$ OR $[x = 0, y \in \{-9,-8.9,\ldots,9\}]$ OR $[x \in \{-9,-8.9,\ldots,9\}, y = -x]\}$ (which are grids along the boundaries of the null spaces for the null hypotheses in $\mathcal{H}$), $J = 3$ power constraints, and 68,003 equality constraints in $\mathbf{A^{(2)}}$. In step two, we used the dual solution to this linear program to approximately identify the active constraints, and then constructed a finer discretization with 9,208,283 variables, 27 Type I error constraints (concentrated in a small neighborhood of the active constraints from the solution to the first step), $J = 3$ power constraints, and 1,314,989 equality constraints in $\mathbf{A^{(2)}}$. The three active Type I error constraints in the optimal solution to the second step correspond to the following pairs of non-centrality parameters $(\Delta_1\{n/(8\sigma^2)\}^{1/2}, \Delta_2\{n/(8\sigma^2)\}^{1/2})$: $(0,0), (0,2.28), (2.28,0)$. This shows that the method of Hampson and Jennison (2014) would not generally work if applied in our setting (which differs from their problem setting). This is because their approach requires that the optimal solution when only constraining Type I error at the global null hypothesis (i.e., non-centrality parameters equal to $(0,0)$) also controls familywise Type I error at all other alternatives.

To solve each sparse linear programming problem, we used the IBM CPLEX solver, version 12.4. To take advantage of the extreme sparse structure of the problem, we used an interior point algorithm. To guarantee high precision, we set the tolerance of the relative duality gap to be 10^{-10}. The linear program in the first step above (coarse version) took approximately 45 minutes to solve, while the second step (finer version) took approximately 12 hours.

Acknowledgment

This work used IBM CPLEX software that was generously made available through the IBM Academic Initiative. This publication's contents are solely the responsibility of the authors and do not represent the views of IBM.

Supplementary Material

Supplementary Material is available on request from the corresponding author. (See email address on first page.)

References

Bauer, P. (1989). Multistage testing with adaptive designs (with discussion). *Biometrie und Informatik in Medizin und Biologie 20*, 130–148.

Bauer, P. and K. Köhne (1994). Evaluations of experiment s with adaptive interim analyses. *Biometrics 50*, 1029–1041.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer.

Boessen, R., F. van der Baan, R. Groenwold, A. Egberts, O. Klungel, D. Grobbee, M. Knol, and K. Roes (2013a). Optimizing trial design in pharmacogenetics research: comparing a fixed parallel group, group sequential, and adaptive selection design on sample size requirements. *Pharmaceutical Statistics*.

Boessen, R., F. van der Baan, R. Groenwold, A. Egberts, O. Klungel, D. Grobbee, M. Knol, and K. Roes (2013b). Optimizing trial design in pharmacogenetics research: comparing

a fixed parallel group, group sequential, and adaptive selection design on sample size requirements. *Pharmaceutical Statistics*.

Brannath, W., E. Zuber, M. Branson, F. Bretz, P. Gallo, M. Posch, and A. Racine-Poon (2009a). Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine 28*(10), 1445–1463.

Brannath, W., E. Zuber, M. Branson, F. Bretz, P. Gallo, M. Posch, and A. Racine-Poon (2009b). Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine 28*(10), 1445–1463.

Bretz, F., H. Schmidli, F. König, A. Racine, and W. Maurer (2006). Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal 48*(4), 623–634.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association 50*(272), 1096–1121.

FDA (2010). Draft guidance for industry. Adaptive design clinical trials for drugs and biologics. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf.

Follmann, D. (1997). Adaptively changing subgroup proportions in clinical trials. *Statistica Sinica 7*, 1085–1102.

Freidlin, B. and R. Simon (2005). Adaptive signature design: An adaptive clinicaltrial design for generating and prospectivelytesting agene expression signature for sensitive patients. *Clin Cancer Res 11*, 7872–7878.

Friede, T., N. Parsons, and N. Stallard (2012). A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in medicine 31*(30), 4309–4320.

Hampson, L. V. and C. Jennison (2014). Optimizing the data combination rule for seamless phaseii/iii clinical trials. *Statistics in Medicine*.

Hochberg, Y. and A. C. Tamhane (1987). *Multiple Comparison Procedures*. New York: Wiley Interscience.

Jenkins, M., A. Stone, and C. Jennison (2011a). An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics 10*(4), 347–356.

Jenkins, M., A. Stone, and C. Jennison (2011b). An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics 10*(4), 347–356.

Jennison, C. and B. W. Turnbull (2007). Adaptive seamless designs: Selection and prospective testing of hypotheses. *J. Biopharmaceutical Statistics*, 1135–1161, doi: 10.1080/10543400701645215.

Lai, T. L., P. W. Lavori, and O. Y.-W. Liao (2014). Adaptive choice of patient subgroup for comparing two treatments. *Contemporary clinical trials*.

Lehmacher, W. and G. Wassmer (1999). Adaptive sample size calculations in group sequential trials. *Biometrics 55*(4), 1286–1290.

Lehmann, E. L. and H. Scheffé (1950). Completeness, similar regions, and unbiased estimation. i. *Sankhyā 10*(4), 305–340.

Marcus, R., E. Peritz, and K. R. Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika 63*, 655–660.

O'Brien, P. and T. Fleming (1979). A multiple testing procedure for clinical trials. *Biometrics 35*, 549–556.

Rosenblum, M., H. Liu, and E.-H. Yen (2014). Optimal tests of treatment effects for the overall population and two subpopulations in randomized trials, using sparse linear programming. *Journal of the American Statistical Association 109*(507), 1216–1228.

Rosenblum, M. and M. J. van der Laan (2011). Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika 98*(4), 845–860.

Russek-Cohen, E. and R. M. Simon (1997). Evaluating treatments when a gender by treatment interaction may exist. *Statistics in Medicine 16*, 455–464.

Schmidli, H., F. Bretz, A. Racine, and W. Maurer (2006). Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: applications and practical considerations. *Biometrical Journal 48*(4), 635–643.

Stallard, N., T. Hamborg, N. Parsons, and T. Friede (2014). Adaptive designs for confirmatory clinical trials with subgroup selection. *Journal of biopharmaceutical statistics 24*(1), 168–187.

Wang, S. J., H. Hung, and R. T. O'Neill (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal 51*, 358–374.

Wang, S. J., R. T. O'Neill, and H. Hung (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subsets. *Pharmaceut. Statist. 6*, 227–244.